

Investigation on Technical Feasibility of FEC Architecture with 1X400Gbps or 4X100Gbps

Xinyuan Wang, Tongtong Wang, Wenbin Yang

Introduction and Background

- This presentation investigates the FEC architecture for 1X400Gbps versus 4X100Gbps implementation based on KP4 RS FEC
- How to stripe ingress data flow to FEC instance is still key item to be investigated for moving 400GbE standard forward
- How to implementation 1X400Gbps RS FEC need to be further investigated

BTI Progress

Item	Proposal	BTI Actions	March Presentations	BTI out of March Plenary
PCS	gustlin_3bs_02_0115	Slide 11 per gustlin_3bs_02 are work items PMD selection influence PCS and FEC Need burst error nature to select PCS and FEC Error model by PMD type?	gustlin_3bs_02_0315.pdf gustlin_3bs_01_0315.pdf	* AM details
FEC	Reference: wang_x_3bs_01_0115	PMD selection BERin required by PMD Try to eliminate unacceptable FEC options e.g. in wang_x_3bs_01 4x100G or 1x400G FEC striping Impact of overspeed on PMD error rates	gustlin_3bs_02_0315.pdf wang_x_3bs_01_0315.pdf	* 4x100 vs. 1x400 decision
PMA	Reference: Slavick_3bs_01_0115 Wang_t_3bs_01_0115 Gustlin_3bs_02_0115	PMD selection and electrical interfaces will impact Muxing scheme	gustlin_3bs_02_0315.pdf wang_t_3bs_01_0315.pdf	* muxing choice related to FEC arch (4x100 vs. 1x400)
Arch	dambrosia_3bs_02b_0115	None	None	
EEE	marris_3bs_01_0115	None	None	
OTN	trowbridge_3bs_01a_0115	none	trowbridge_3bs_01_0315.pdf	* reuse of modules (depends on PMA muxing)

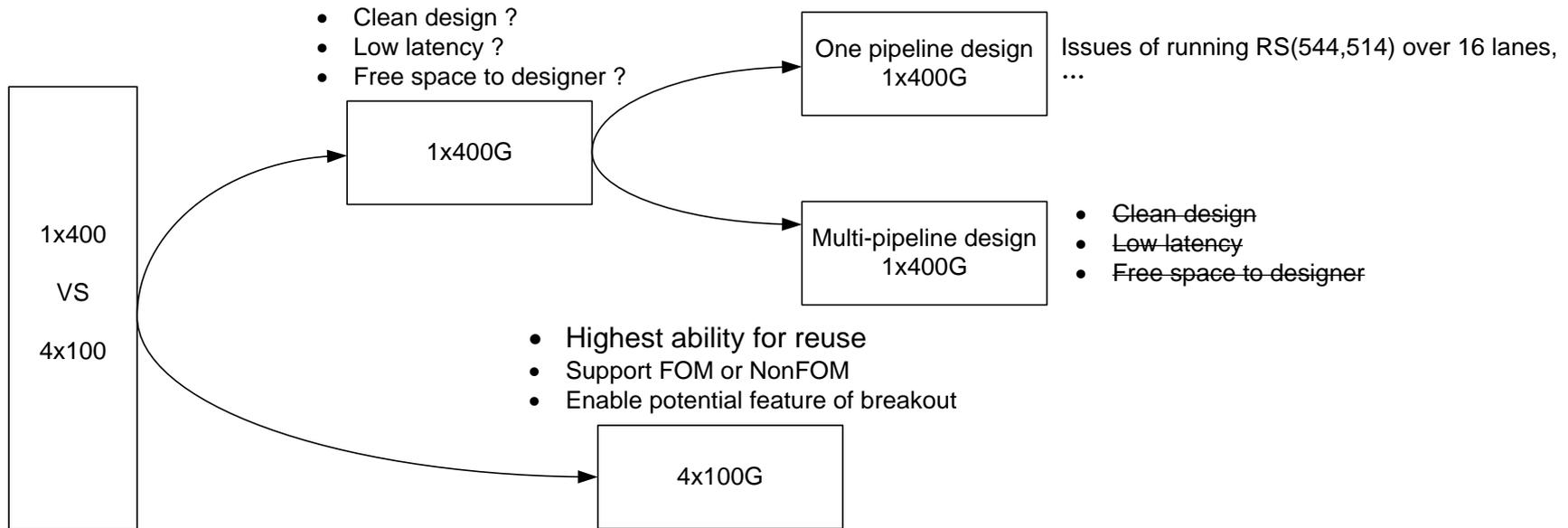
[bti_3bs_01_0315](#)

[wang_x_3bs_01_0315](#)

Issues in Implementing 1X400Gbps RS(544,514)

- Not a straight forward evolution from mature 100Gbps KP4 FEC and will impact on 400GbE architecture
 - › Distribution over 16 lanes instead of 4 lanes. Common design of data bus width for 1x400Gbps RS(544,514) FEC is 680bit (NOT 640bit) to finish FEC codeword in 8 cycles, and this 680bit data bus is not divisible for 16 lane.
 - › 100G KP4 FEC works because it distributes FEC codeword on 4 lanes, as described in [wang_x_3bs_01a_0115](#) FEC choice rule 4, "FEC block size (n*m) should be divisible by (lane_number*m), for example, 4 lanes in 100G KR4/KP4 FEC"
- Solutions base on current process technology
 - › Option 1: FEC function running at 680bit@~664MHz, use 680/640b gearbox for fitting in Serdes interface
 - › Option 2: ~10% Over-clocking from 680bit@~664MHz into 640bit@~730MHz with extra pad
 - › Both options are not clean/straightforward design, can we avoid it from architectural design?
- More problem – AM spacing complexity
 - › Even if FEC block manage to fit 640bit, it takes 8.5 cycles for RS(544,514) codeword. How to guarantee AM header in FEC codeword is placed on 16 lanes properly?

400GbE FEC Architecture Exploration

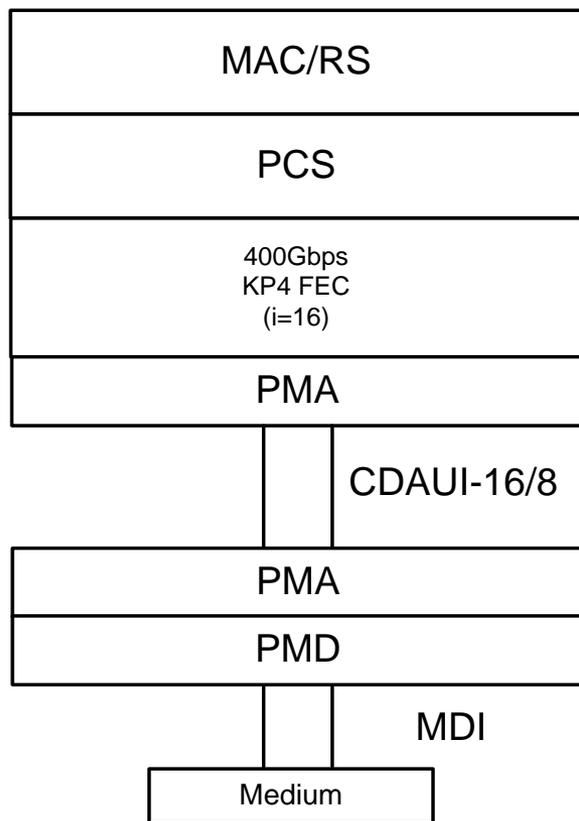


□ Current observations:

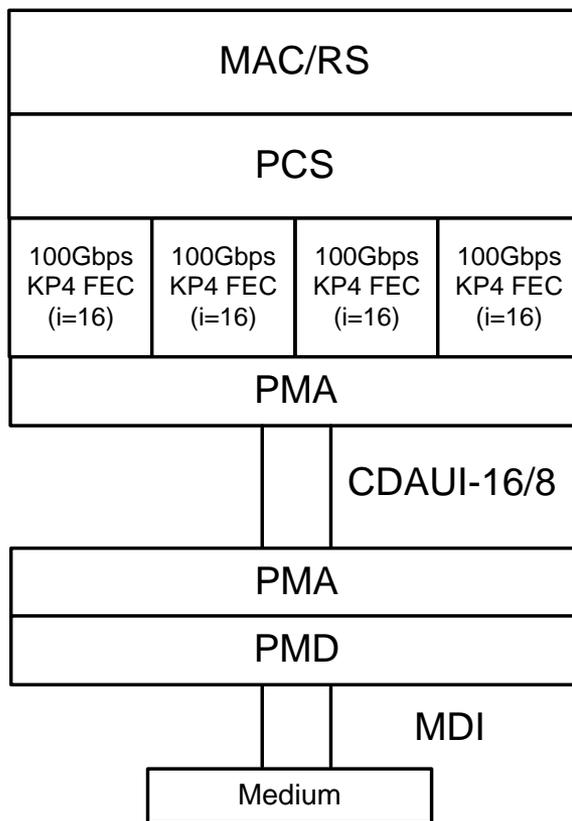
- Try to find lowest latency and cleanest FEC architecture for 400GbE project
- Having issues in implementing one pipeline RS(544,514) over 16 lane. it is NOT a clean and lowest latency choice to us now
- Multi-pipeline 1x400G FEC (a.k.a 4x100G TDM to form a 1x400G bps FEC) is also not a good backup option due to cost and complexity
- 4x100G FEC can support FOM or NonFOM option, and has highest ability for reuse. Potential merits in enabling break out feature.

400GbE FEC Architecture Options

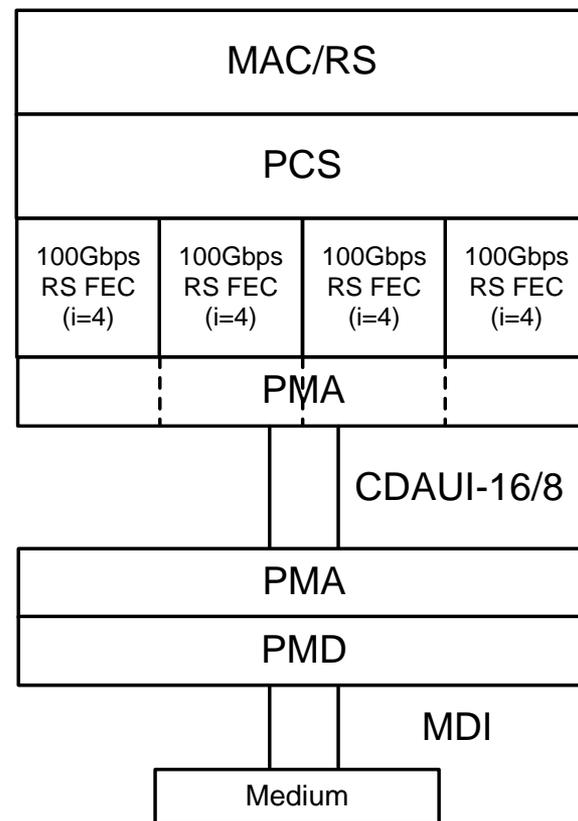
- Arch A: one pipeline
1X400G FEC



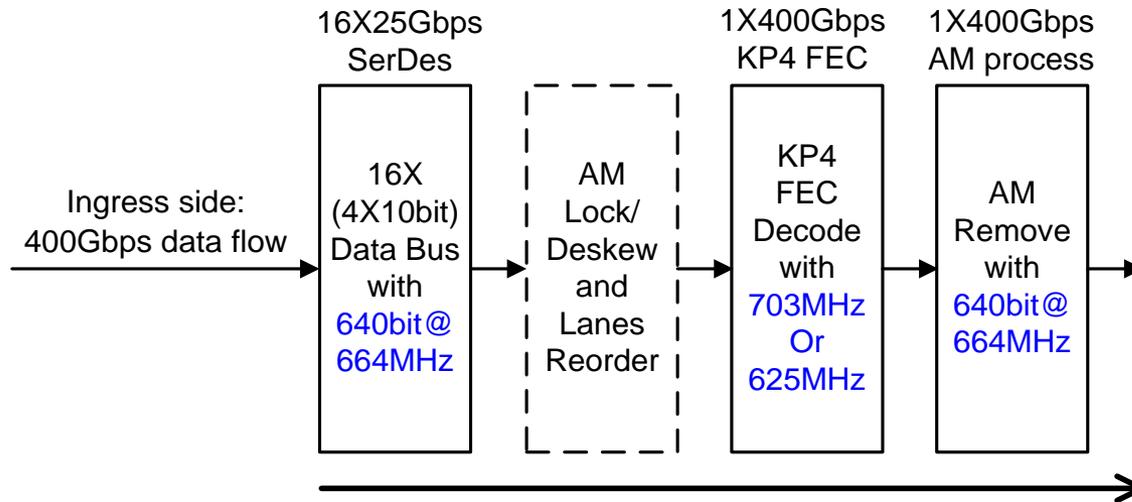
- Arch B: Multi-Pipeline
4X100G FEC to form a
1x400G black box



- Arch C: 4X100G FEC



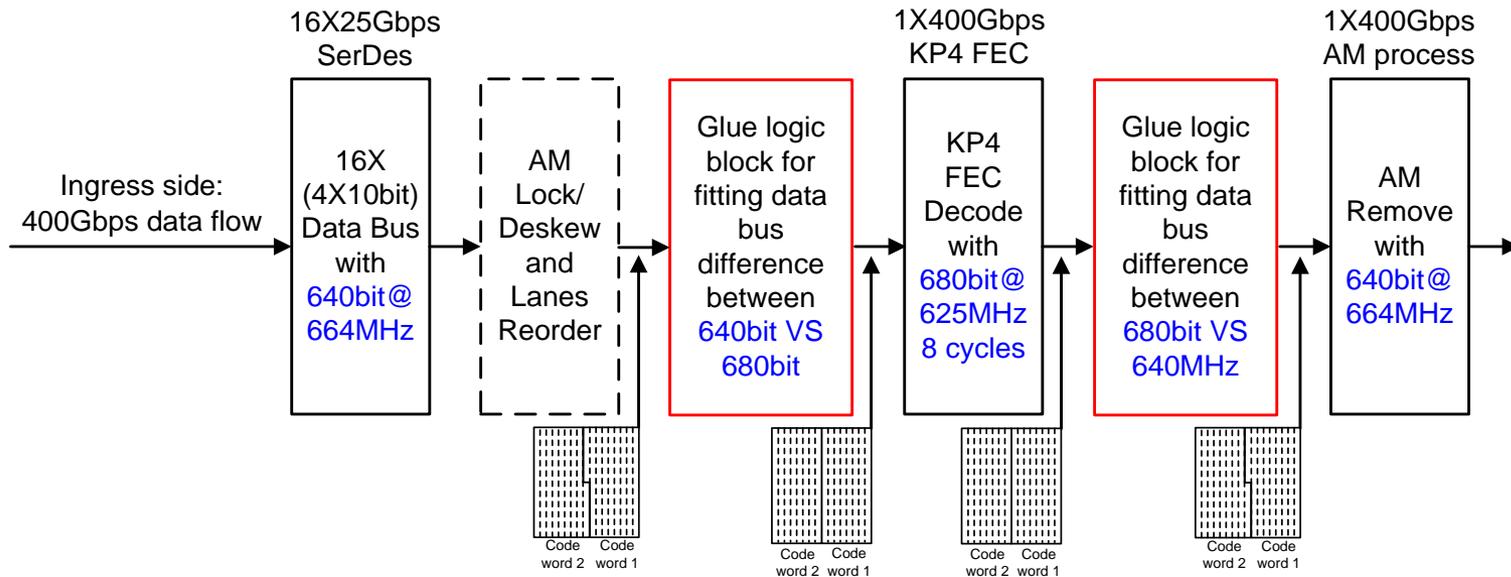
Issue for Arch A (1X400G RS(544,514)) over 16 lanes



- A simple issue is that $544/64 (= 8.5)$ is not an integer*
 - That means, physically, current KP4 FEC design need to be re-considered to work with offset, more cost(area/latency) needed to adopting current SerDes interface
 - Option 1: running at 680bit@625MHz data bus in 8 cycle to complete one FEC codeword encode/decode
 - Option 2: running at 640bit@703MHz data bus in 9 cycle to complete one FEC codeword (over clocking)
 - Option 3: running at 640bit@625MHz data bus in 8 cycle, more logic inside RSFEC block to process the offset
- AM header must be distributed and restored traversing 16 Lanes, thus 160bit granularity is mandatory and higher complexity in option 1 due to data bus width mismatch in function block

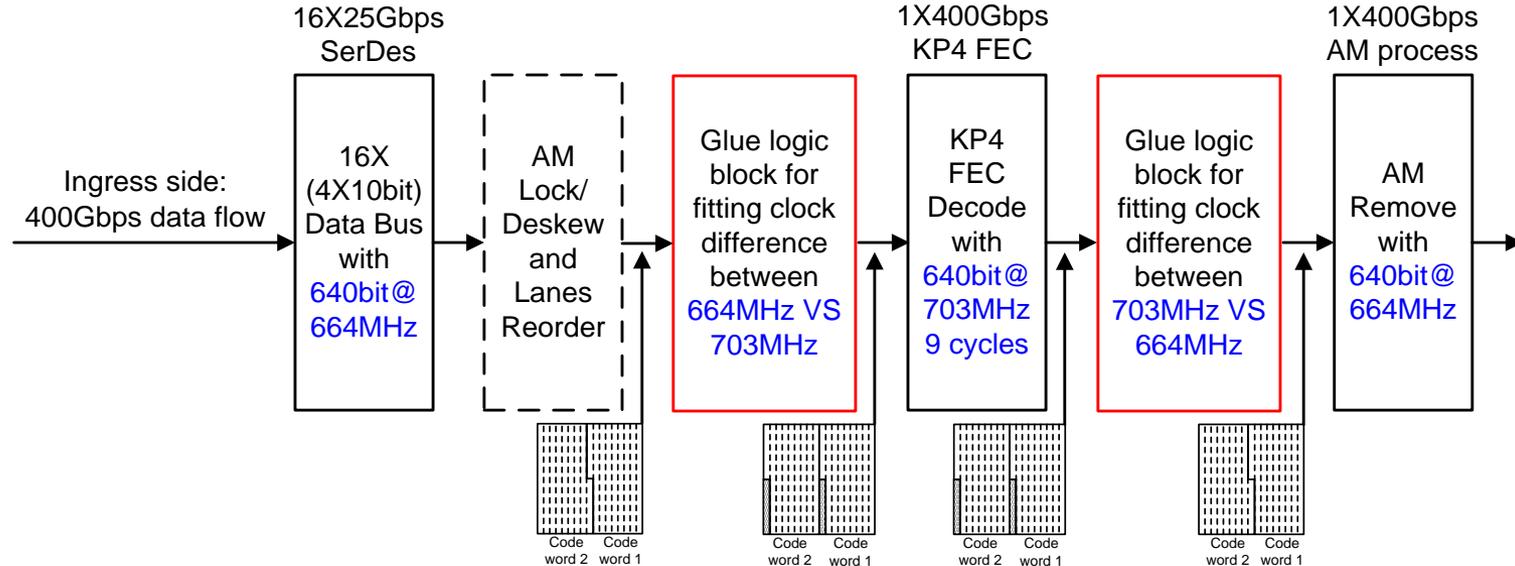
*For 100G .bj FEC over 4 lanes, $544/16 = 34$

Implementation of Option 1 of Arch A:



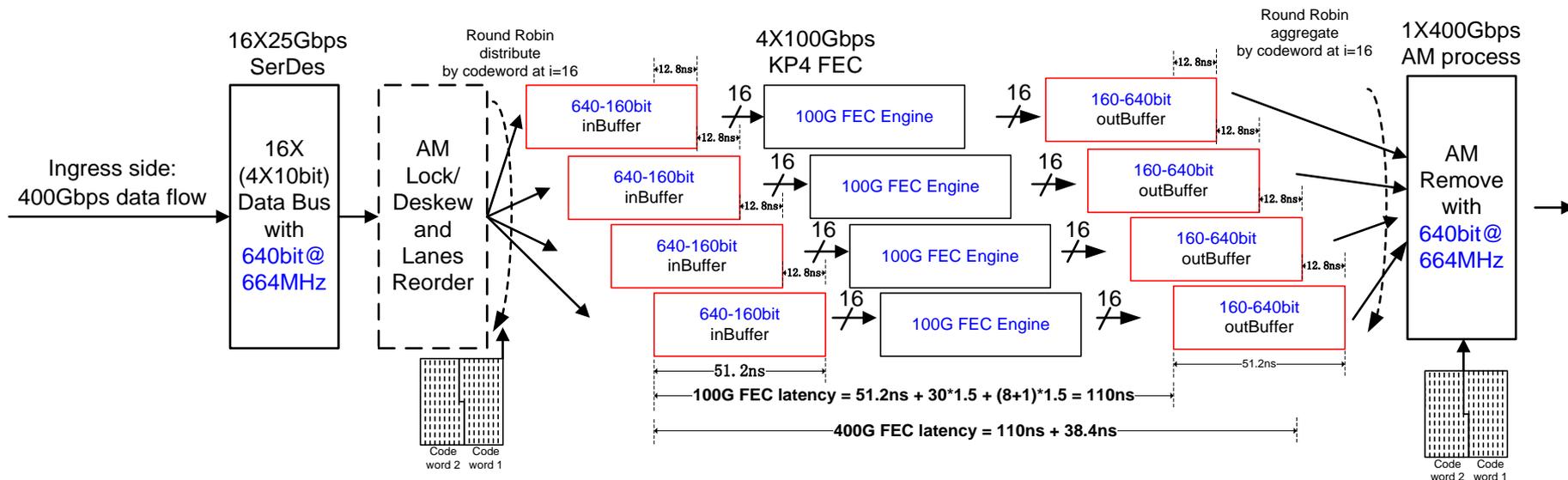
- Considering ingress example as in above diagram, data streams come from 16x26.5625G SerDes and need to do AM lock/de-skew/reordering before FEC decoder with “Offset” at data bus width
- For option 1:
 - Decode FEC codeword in 8 cycles @680bit@625MHz
- Additional logic in RED block needed for fixing this data bus width difference

Implementation of Option 2 of Arch A:



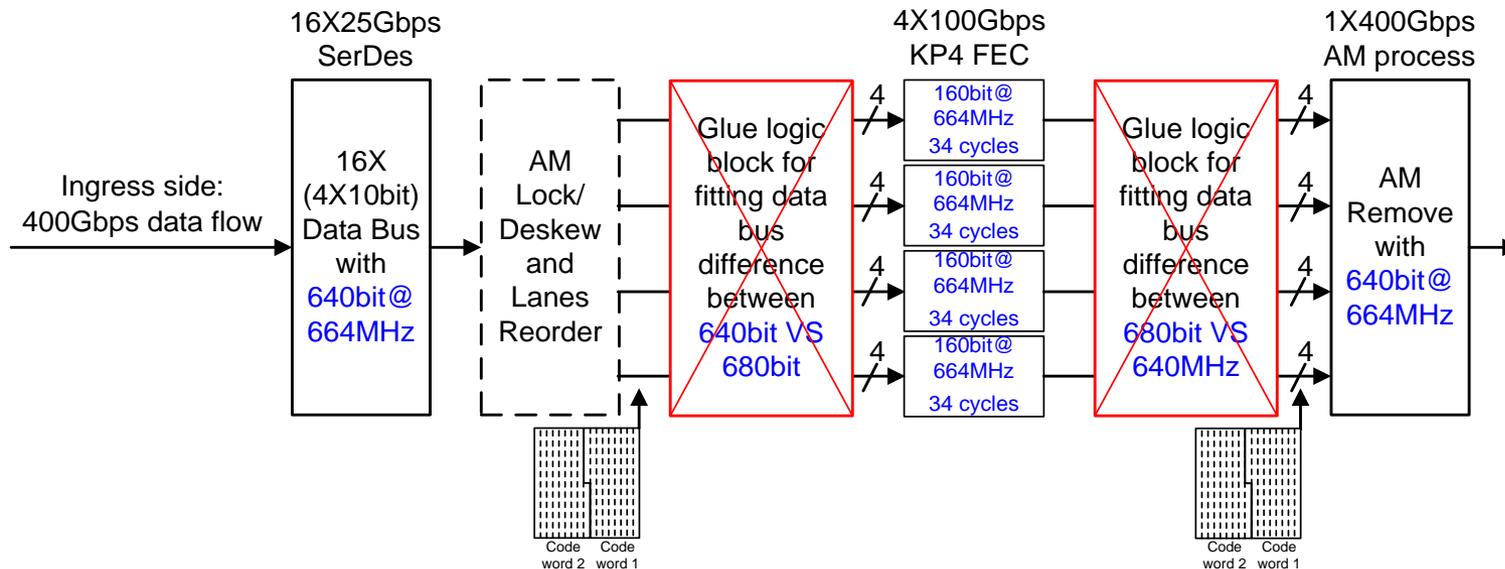
- Considering receiver example as in above diagram, data streams come from 16x26.5625G SerDes and need to do AM lock/de-skew/reordering before FEC decoder with “Offset” at clock rate
- For option 2:
 - Decode FEC codeword in 9 cycles @640bit@703MHz with pad included
- Additional logic in RED block needed for fixing this clock rate difference

Issues for Arch B: Multi-pipeline 1x400G RS(544,514) by 4X100G FEC TDM



- ❑ 400G data flow distributes to 4 100G FEC engines by round robin
- ❑ Need two sets of codeword buffers for fitting data rate gap between 400Gbps and 100Gbps. Each inBuffer fills up in at least 12.8ns and empty in 51.2ns, and outBuffer fills up in 51.2ns and drains in 12.8ns.
- ❑ 100G KP4 FEC latency = $51.2\text{ns} + (30+8+1)\text{cycle}@664\text{Mhz} \approx 110\text{ns}$
- ❑ 400G FEC latency on Arch B $\approx (110\text{ns} + 38.4\text{ns})$
- ❑ This architecture has no latency or area advantages

Arch C: 4X100G FEC



- Same data width and clock rate for SerDes/FEC/AM block at **34 cycles @4X(160bit)@664MHz** is fully compatible to 802.3bj KP4 FEC without additional distribution and aggregate logic cost and buffering delay
- 5440bit codeword distributes to 4 lanes rather than 16 lanes as in Arch A/B implementation. No additional glue logic or multiple clock domain required
- This is straight forward evolution from mature 802.3bj KP4 FEC design

Perspective of Future Process Technology

- The table below shows all possible implementations at different clock rate/process node for 400Gbps KP4 FEC on 1x400Gbps (Arch A) and 4x100Gbps (Arch C)
- 4X100Gbps KP4 FEC is much simpler because it can finish in integer clock cycles for 160bit@664MHz or 320bit@332MHz, which are the most popular designs in current ASIC and FPGA technology
- Next good clock rate option for both 400G and 100G KP4 FEC is on 1.328Ghz, how far away is that?
- Another issue: >1000bit data bus width in 400Gbps FEC implementation is in higher risk in wiring/timing convergence

Numer of Symbols	Data Bus Width Per Lanes(Bit)	Clock Rate	Data Bus Width Per 400Gbps FEC(Bit)	Number of clock cycle for 400Gbps	Data Bus Width Per 100Gbps FEC(Bit)	Number of clock cycle for 100Gbps
1	10	2.65625GHz	160	34	40	136
2	20	1.328GHz	320	17	80	68
3	30	885MHz	480	11.333	120	45.333
4	40	664MHz	640	8.5	160	34
5	50	531MHz	800	6.8	200	27.2
6	60	443MHz	960	5.667	240	22.667
7	70	379MHz	1120	4.857	280	19.429
8	80	332MHz	1280	4.25	320	17
9	90	295MHz	1440	3.778	360	15.111
10	100	265MHz	1600	3.4	400	13.6

Comparison of 400GbE FEC Architecture

	Pros	Cons
Arch A: 1X400Gbps one-pipeline KP4 FEC	One FEC instances Architecture	Additional logic complexity in current and near the future process technology
	Possibly lower latency at ~75ns@664MHz with one-pipeline design	Additional 4X100GbE FEC logic if enable breakout into 4X100GbE
		Re-development 1X400Gbps KP4 IP Core
		Non-compatible logic/hardware solution in 1X400GbE and 4X100GbE interface
		Risk on wiring/time convergency
		Assume bit mux in PMA: Limited BER in electrical interface with burst error
		Assume bit mux in PMA: No burst error tolerate in optical link generally
		Rule out 4 100Gbps instance gearbox with FEC integrated
Arch C: 4X100Gbps KP4 FEC		Rule out FOM for muxing in PMA
	No glue logic required for fitting data width/clok difference in real implementation	Multi-FEC instances Architecture
	Unfied architecture to Enable breakout into 4X100GbE	Possibly longer latency:~110ns@664MHz
	Max reuse 802.3bj KP4 IP Core	
	Unified logic layer solution for 1X400GbE and 4X100GbE interface	
	Easier on wiring/timing convergency	
	Assume bit mux in PMA: more relax in hardware design on electrical link with burst error	
	Assume bit mux in PMA: more robust and FEC performance in optical link with burst error	
	Enable 4 100Gbps instance gearbox with FEC integrated	
Support FOM or Non-FOM bit mux in PMA		

From “The Architecture is a Deliverable”

Implicit Objective

- An architecture is implicit and not stated in the objectives
- But- it is a first-class output of the standard.
 - It frequently has a life beyond the original project
 - It can enable electrical interface evolution
 - It can enable future IEEE and non-IEEE PMDs
 - It can enable necessary system partitioning
- This helps the broad market potential

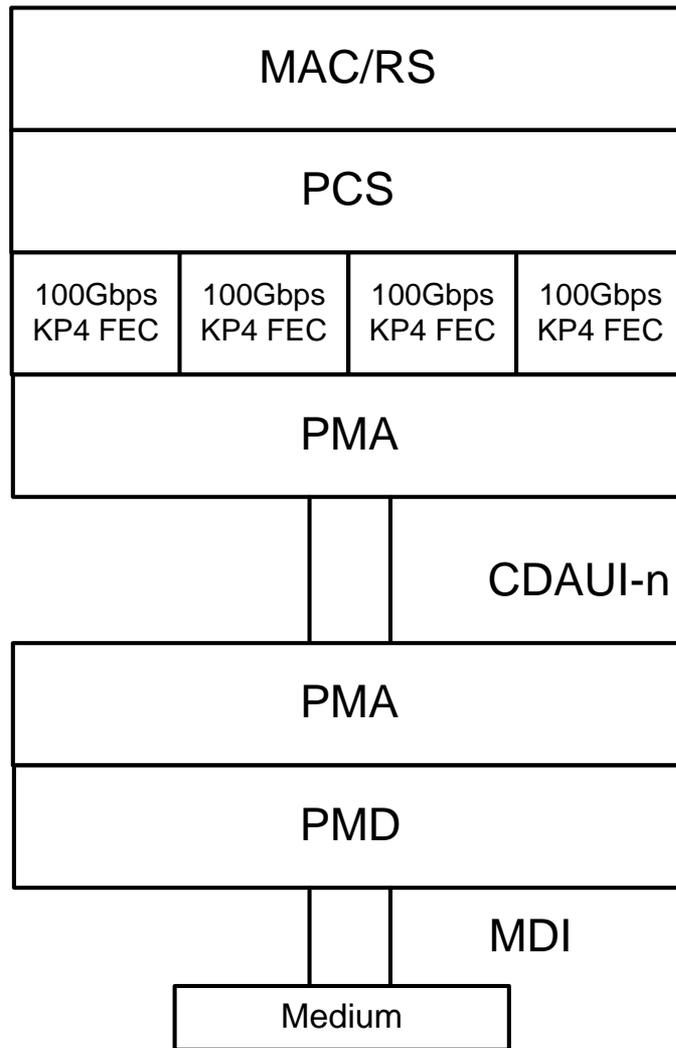
Life beyond original standard

- A successful architecture can support future PMDs and modules
 - IEEE, proprietary, and/or MSA-based

[ofelt 3bs 01a 0115](#)

- ▣ Arch C with 4X100Gbps FEC proposal is significantly more robust architecture right now and in near future

Proposal for 400GbE Logic Layer with KP4 FEC



- 4X100Gbps KP4 FEC Parallelism
- Support either FOM or Non-FOM bit Mux in PMA
- Support FOM interoperation with Non-FOM implementation and vice versa on non-bursty links.

Summary

- The FEC architecture proposal with 4X100Gbps FEC in parallel is a more simple solution, it will not only lower total area cost in 400GbE & 4X100GbE compatible design and also enable breakout feature, reuse IP cores and unified line card design and lead to broader market potential
- The FEC architecture proposal with 4X100Gbps FEC in parallel is a more robust system, which can provide maximum coding gain to optical link and achieve better performance in the face of burst errors. It will enable diverse PMD solutions that are not limited by current 802.3bs objectives

Thank you